Laboratorium 2. Wprowadzenie do Pythona

Projektując środowisko eksperymentalne z metodami przetwarzania języka naturalnego jednym z pierwszych kroków jest przygotowanie odpowiedniego zestawu danych. W tym przypadku oznacza to pewne repozytorium treści zapisanych w języku naturalnym. Laboratorium to pozwoli poznać techniki na pozyskiwanie treści z wybranego serwisu informacyjnego. Rozwiązaniem tego laboratorium będzie bardzo przydatne narzędzie do pozyskiwania danych.

Zadanie 2.1. Scrap

Pierwsze zadanie to implementacja programu, który scrapuje wskazane dane z lokalnego serwisu informacyjnego Gazeta Wrocławska - Wiadomości[1]

1. Zainstaluj bibliotekę Scrapy i utwórz startowy projekt
2. Zaimplementuj program, który pobiera tytuł dowolnego artykułu z serwisu "Gazeta Wrocławska - Wiadomości"
3. Użyj do tego selektora `XPath` oraz funkcję `Inspect` w przeglądarce
4. Zapoznaj się z plikiem `robots.txt` serwisu "Gazeta Wrocławska - Wiadomości"
5. W skrypcie pod zadaniem napisz komentarz wyjaśniający znaczenie oraz zawartość tego pliku

Dokumentacja:

- `Scrapy installation`[2]
- `Scrapy introduction`[3]
- `XPath selector`[4]
- `Inspect`[5]

Zadanie 2.2. Crawl

Kolejne zadnie to implementacja programu, który pobiera dane z więcej niż jednej strony automatycznie szukając nowych hiperłączy.

1. Napisz program typu Crawler:
   - Program powinien odnajdywać na stronie głównej hiperłącza do artykułów
   - Następnie dla każdego znalezionego hiperłącza, które należy do serwisu, pobierz tytuł artykułu
   - Ustaw opóźnienie na co najmniej 2 sekundy
   - Rozwiązanie może być rozszerzeniem skryptu z poprzedniego zadania
2. Pobrane tytuły wypisz na terminalu
3. W skrypcie pod zadaniem napisz komentarz wyjaśniający w jakim celu ustawia się opóźnienie

Dokumentacja:

- `Crawl Spider`[6]

Zadanie 2.3. Scrap and Crawl

Ostatnie zadanie ma na celu rozwinąć dwa poprzednie zadania o pobieranie większej ilości danych z jednej strony oraz zapis tych danych do pliku.

1. Rozszerz skrypt z poprzedniego zadania tak, aby pobierać:
   - Tytuł (zaimplementowane w pierwszym zadaniu)
   - Data publikacji
   - Tagi
   - Autor lub autorzy
   - Treść artykułu
2. Pobrane dane zapisz do pliku JSON:
   - Jako nazwę pliku użyj daty publikacji oraz tytuł artykułu (zamień spacje na znak podkreślenia)
   - Do zapisu danych wykorzystaj słownik oraz `json.dump()`
3. Pobierz co najmniej dane z co najmniej 60 artykułów.

Dokumentacja:

- `json.dump()`[7]

[1] <https://gazetawroclawska.pl/wiadomosci/>
[2] <https://docs.scrapy.org/en/latest/intro/install.html>
[3] <https://docs.scrapy.org/en/latest/intro/tutorial.html>
[4] <https://docs.scrapy.org/en/latest/topics/selectors.html#working-with-xpaths>
[5] <https://blog.hubspot.com/website/how-to-inspect>
[6] <https://docs.scrapy.org/en/latest/topics/spiders.html?highlight=rule#crawlspider>
[7] <https://docs.python.org/3/library/json.html>